COURSE: DSA-5103 – INTELLIGENT DATA ANALYTICS

SECTION: 001

SEMESTER: FALL 2022

INSTRUCTOR: DR. CHARLES NICHOLSON

TITLE: COURSE PROJECT REPORT

NAME OF PROJECT: DISEASE PREDICTION

GROUP NUMBER: 18

GROUP MEMBERS:

Varshitha Vasireddy – varshitha.c.vasireddy@ou.edu

Biswas Nandamuri – biswas.nandamuri@ou.edu

Vasu Dev – vasu.janapala@ou.edu

Pranav Vichare – pranav.b.vichare-1@ou.edu

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1. Problem statement and objective

Predictive Analytics can identify specific risk factors for various populations. For example, it can identify patients with the highest probability of hospitalization. Also, if it is an infectious disease, it can help us prevent a significant crisis. The existing statistics say that data analytics will be crucial soon for the healthcare industry, and it is becoming very very crucial in clinical, operational, and financial sectors.

Disease Prediction plays a pivotal role in healthcare informatics. The symptoms of a patient can identify most diseases. Multiple Diseases can have the same symptoms, so doctors generally suggest lab tests to identify the disease correctly. However, lab tests are costly and time-consuming. If identifying a disease based only on the symptoms is achieved, then appropriate lab tests can be suggested to the patient. This helps identify the disease swiftly, and as a result, treatment can be provided immediately.

The main aim of this project is to use machine learning methods to predict the top 3 diseases based on symptoms displayed by the patient. The dataset is taken from Kaggle, consisting of 41 unique diseases with 131 symptoms.

We then analyzed different and parallel classification systems for disease prediction as it enhances the computational efficiency of results. Feature selection decreases the accuracy of the model; hence all symptoms are considered for modeling. A pdf report is generated with the top 3 diseases, their probabilities, and a short description of the disease.

2. Data Understanding

2.1. Data description

The dataset consists of diseases and their symptoms. There are 18 columns in the dataset, the disease column is the first one, and the rest columns are symptoms. Altogether, it has 4920 observations and 18 variables. If the value under the disease attribute has more than a word, it splits with a space, whereas under the symptom columns, a value with more than a word is separated with an underscore, and also in the same columns, data has leading, trailing, and between spaces in it. Also, the symptom columns have missing data, which are represented using NAs. The symptoms are not unique values for diseases.

This data is referred to as raw data. A few rows and columns of raw data are as below.

Table 2.1: Few rows and columns of the raw dataset

Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5
Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches	
Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches		
Chronic cholestasis	itching	vomiting	yellowish_skin	nausea	loss_of_appetite

Chronic cholestasis	vomiting	yellowish_skin	nausea	loss_of_appetite	abdominal_pain
Chronic cholestasis	itching	yellowish_skin	nausea	loss_of_appetite	abdominal_pain
Diabetes	fatigue	weight_loss	restlessness	lethargy	irregular_sugar_level
Diabetes	fatigue	weight_loss	restlessness	lethargy	irregular_sugar_level
Diabetes	weight_loss	restlessness	lethargy	irregular_sugar_level	blurred_and_distorted_vision

Another dataset is used, which consists of description about all the 41 diseases that are present in raw data. Only two columns are present in this dataset, where one is disease name, and another is description.

Table 2.2: Few rows and columns of disease description dataset

Disease	**Description**
Diabetes	Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your primary source of energy and comes from the food you eat. Insulin, a hormone made by the pancreas, helps glucose from food get into your cells to be used for energy.
Malaria	An infectious disease caused by protozoan parasites from the Plasmodium family that can be transmitted by the bite of the Anopheles mosquito or by a contaminated needle or transfusion. Falciparum malaria is the most deadly type.

These datasets are taken from kaggle from the link: https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset?select=dataset.csv. Few changes are made to the dataset of disease description manually to match the disease names present in both the datasets.

2.2. Data Preprocessing

Raw data consists of Diseases and the 17 Symptoms that the patient reported. It is in long form. It is to be preprocessed to get the desired and cleaned dataset. No preprocessing is done on the symptom description dataset.

Data preprocessing steps

- Initially, when reading raw data into a data frame, the empty cells are replaced as NAs.
- Then, all the symptom columns are trimmed of leading or trailing spaces. In the dataset, a symptom with more than two words is already joined with an underscore, and an extra space between them is removed.
- This processing helps in removing all the spaces in symptom columns, as a result making the data clean and consistent.
- NA cells of symptom columns are replaced with a -1 value.
- Now all symptoms in the raw data are converted into dummy variable representations by
    - Extracting all symptoms in the raw dataset into a vector
    - Then create a new data frame with all the diseases in the first column and the extracted symptoms as the column names

- Lastly, the cells corresponding to a disease and a symptom are marked as TRUE or FALSE depending on the raw data

- This preprocessing is done so that model doesn't misinterpret the same symptom on two different symptom columns (Symptom_1, Symptom_2, ..., Symptom_17) as two different symptoms because under the hood, "R" creates dummy variables for all levels of factor variables. So now, unique symptoms are considered during modeling.
- Duplicate rows are also removed.
- Lastly, as there are no more missing values in the dataset, imputation is not needed.

This dataset is called **clean.data** and looks as below

Table 2.2: Few rows and columns of clean data

Disease	abdominal_ **pain**	abnormal_ **menstruation**	acidity	acute_liver_ **failure**	altered_ **sensorium**	anxiety
Fungal infection	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Chronic cholestasis	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
Hyperthyroidism	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
Osteoarthritis	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

So, in the clean data, we have a column with disease names and different symptom names. If that symptom is observed in that disease, then TRUE is assigned as a value, else FALSE is assigned. In total, there are 41 diseases and 131 symptoms.

2.3. Exploratory data analysis

Clean data consists of non-numeric values, i.e., either TRUE or FALSE, and a report is generated

Table 2.3: Few columns of clean data non-numeric data quality report

variable	n	missing	missing_pct	unique	unique_pct
Disease	304	0	0	41	13.49
abdominal_pain	304	0	0	2	0.66
abnormal_menstruation	304	0	0	2	0.66
acidity	304	0	0	2	0.66
acute_liver_failure	304	0	0	2	0.66

The above report shows the variable and the total number of rows in the dataset and the missing values, missing values percentage, unique values, and unique values percentage. After preprocessing the data, we got 304 rows of clean data, and as said above, 41 unique diseases are present, and every symptom is either assigned TRUE or FALSE; hence 2 unique values are present for each symptom.

Visualization 1: Top 5 symptoms of clean data

The top 5 symptoms that are primarily observed in all 41 diseases of clean data are as below.

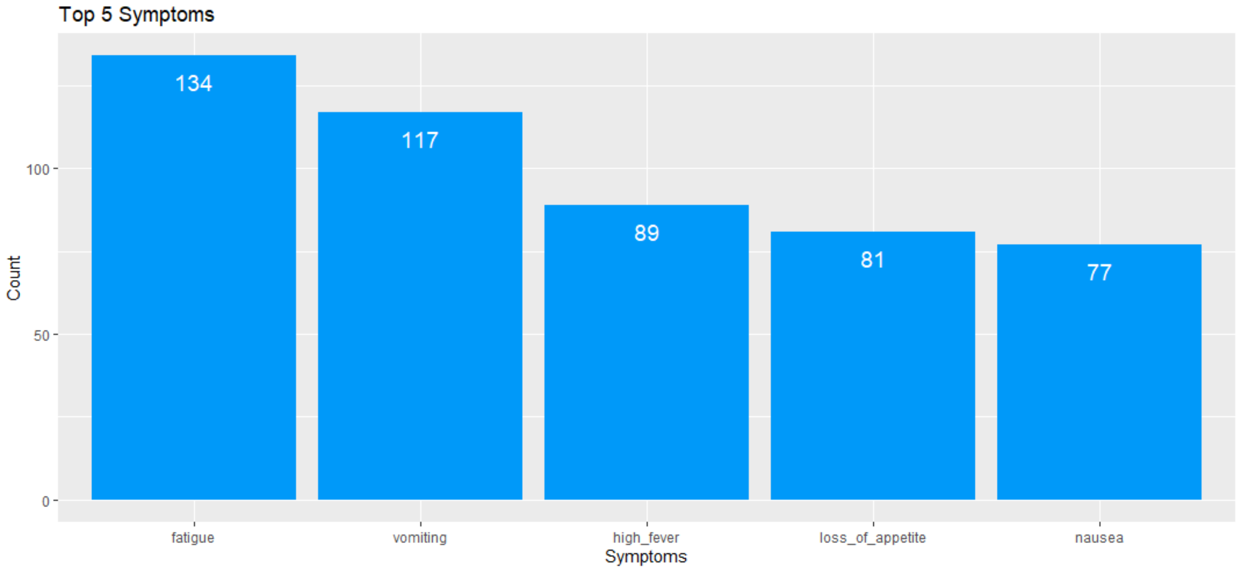


Figure 2.1: Clean data top 5 symptoms among diseases

From the above figure, it can be seen that **Fatigue** is the most common symptom seen among the diseases, with the following 4 being **vomiting, high fever, loss of appetite, and nausea.** So if one gets affected by any disease, there is a high chance of noticing any of these symptoms.

Visualization 2: Correlation matrix between symptoms

If the correlation value between the 2 variables is between 0.90 to 1.0, then they are said to be very highly positively correlated variables. Below is a table with highly positive correlated variables and their correlation value.

Table 2.5: Clean data highly correlated variables correlation values

Var1	**Var2**	**Freq**
acute_liver_failure	coma	0.94
increased_appetite	irregular_sugar_level	0.94
anxiety	palpitations	0.94
drying_and_tingling_lips	palpitations	0.94
irregular_sugar_level	polyuria	0.94
anxiety	slurred_speech	0.94
drying_and_tingling_lips	slurred_speech	0.94
acute_liver_failure	stomach_bleeding	0.94

receiving_blood_transfusion	yellow_urine	0.94
receiving_unsterile_injections	yellow_urine	0.94
abnormal_menstruation	mood_swings	0.94
brittle_nails	cold_hands_and_feets	0.93
cold_hands_and_feets	enlarged_thyroid	0.93
brittle_nails	puffy_face_and_eyes	0.93
enlarged_thyroid	puffy_face_and_eyes	0.93
cold_hands_and_feets	swollen_extremeties	0.93
puffy_face_and_eyes	swollen_extremeties	0.93
brittle_nails	weight_gain	0.93
enlarged_thyroid	weight_gain	0.93
swollen_extremeties	weight_gain	0.93

Below shows the heat map of highly correlated variables

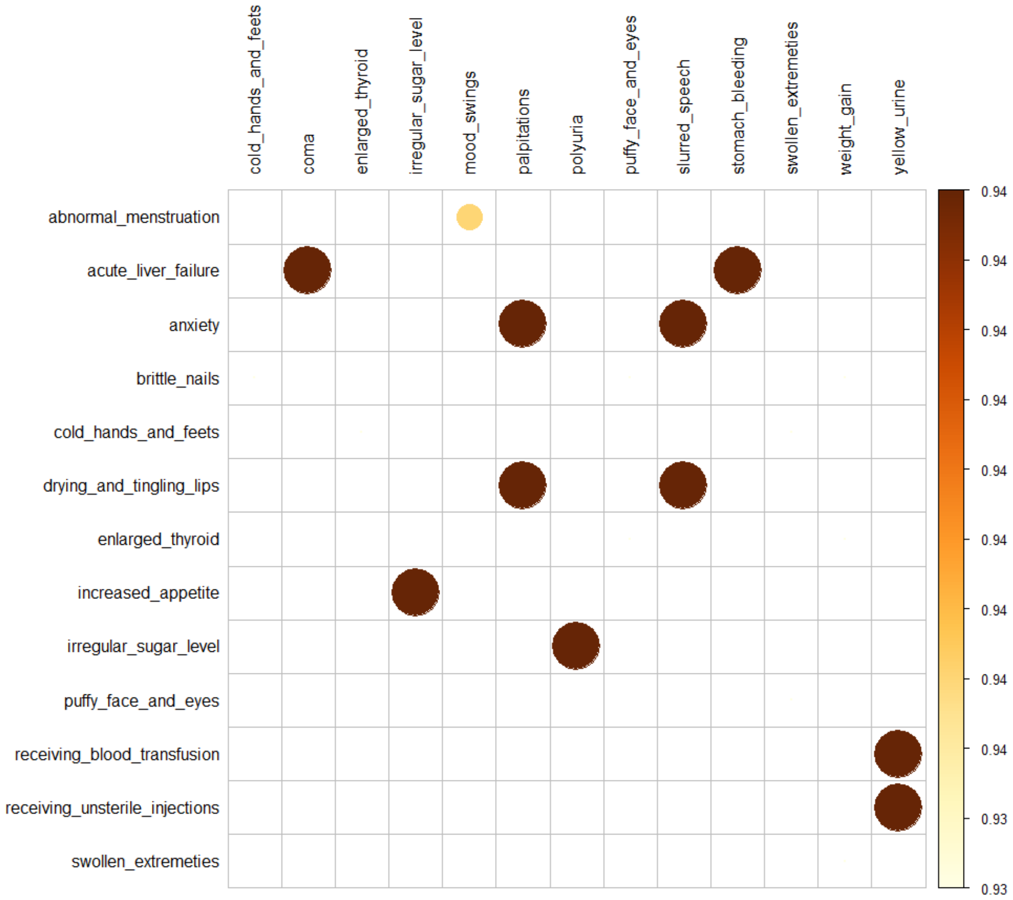


Figure 2.2: Clean data highly correlated variables heat map

3. Methodology

All the variables are essential for predicting the disease so there is no feature selection for this dataset.

For the given data, any machine learning model needs to predict one of the 41 diseases in the disease column. So, considering this to be a classification machine learning problem, we initially used Decision Trees. In terms of the model with respect to the CARET package, we used conditional inference tree (ctree) because ctree uses a statistical approach to identify the most critical symptoms in the data, and then builds the decision tree based on these symptoms.

Upon training the ctree model using the below hyperparameters:

$$mincriterion = [0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00, 1.05]$$

Using 5-fold cross-validation with 3 repeats, the resulting cross-validation training curve is as follows:

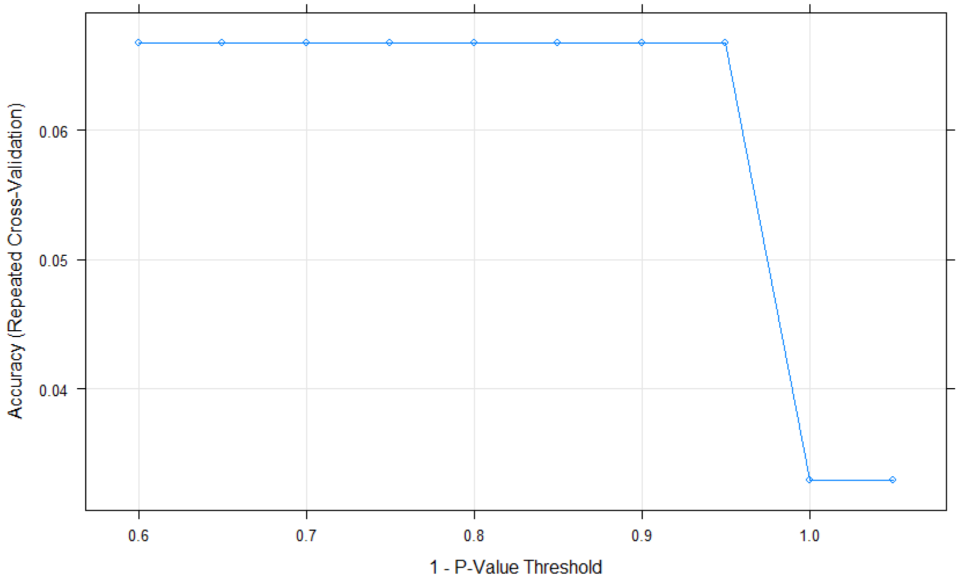


Figure 3.1: 5-fold CV with 3 repeats training curve of decision tree model

The maximum accuracy and kappa values achieved by ctree model is ~6% and ~3.5% at mincriterion of 0.95 value. This ctree model did not perform as well as we thought.

The next machine learning model we considered is an ensemble model called the Random Forest model because we wanted to see if multiple randomly selected symptoms based on multiple sub-trees might return a better result. Using the exact training mechanism of 5-fold cross-validation with 3 repeats and using the following values for Random Forest's hyperparameter:

$$mtry = [60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72]$$

Surprisingly the resulting model yielded accuracy and kappa values of ~99% and ~99%.

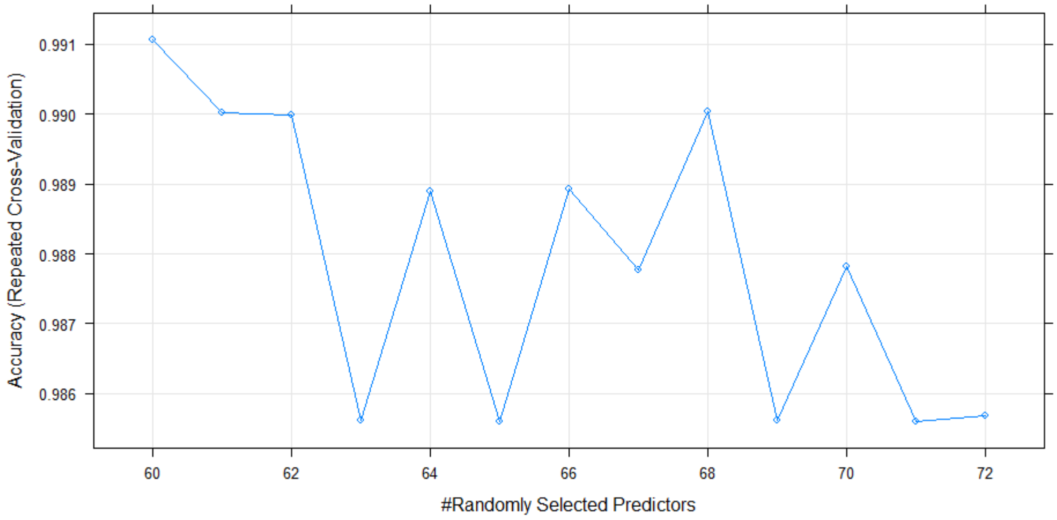


Figure 3.2: Tuning curve generated for the Random Forest Model

Based on the trained Random Forest model, the top 20 important symptoms to predict the disease are as follows:

Table 3.1: The top 20 important symptoms selected by Random Forest Model

muscle_pain	family_history
pain_behind_the_eyes	joint_pain
mild_fever	visual_disturbances
rusty_sputum	stomach_bleeding
red_spots_over_body	weight_loss
dark_urine	receiving_blood_transfusion
blood_in_sputum	nausea
yellowing_of_eyes	slurred_speech
itching	increased_appetite
coma	toxic_look_(typhos)

Below are other performance metrics of the Random Forest Model

Sensitivity is 1 for all classes (i.e., all diseases) except for Hepatitis C and D, which are 0.857 and 0.9, respectively. Sensitivity being 1 tells that the model is predicting the diseases accurately except for two diseases, which is why the model lacks 100% accuracy. Specificity is 1 for all classes except for Chronic Cholestasis, which is 0.99. Specificity being 1 tells that the model is able to distinguish well between different classes of data, except for 1 class, which is also almost 1, but not perfect. The precision score is 1 for all classes except for Chronic Cholestasis, which is 0.8. Precision being 1 tells that positive examples are identified without any errors. The recall score is 1 for all classes except for Hepatitis C and D, which are 0.857 and 0.9, respectively. The recall score being one tells that model is able to identify all positive examples. F1 score is 1 for all classes except for Chronic Cholestasis, Hepatitis C, and D, which are 0.89, 0.92, and 0.947, respectively. The F1 score is one that tells that model is able to make accurate positive predictions while also correctly identifying all of the positive examples in the data, which tells that the model has a perfect balance between precision and recall. From all the performance metrics, it could be seen that the model is able to make accurate predictions with a very high degree of accuracy for the data provided to the model.

Lastly, trying another ensemble model called XGBoost, we wanted to check the difference between Random Forest and XGBoost's methods of training the sub-models, hoping we might get the perfect accuracy and kappa scores of 100%. Thus, using the same method of 5-fold cross-validation with three repeats and using a whole bunch of hyperparameters shown below:

Table 3.2: Hyperparameters of XGBoost

eta	[0.3, 0.4]
max_dept	[1, 2, 3]
colsample_bytree	[0.6, 0.8]
subsample	[0.5, 0.75, 1]
nrounds	[50, 100, 150]

Similar to the above Random Forest model, the XGBoost returned with exact accuracy and kappa values of ~99% and ~99%. But XGBoost took relatively more time to train compared to the Random Forest model.

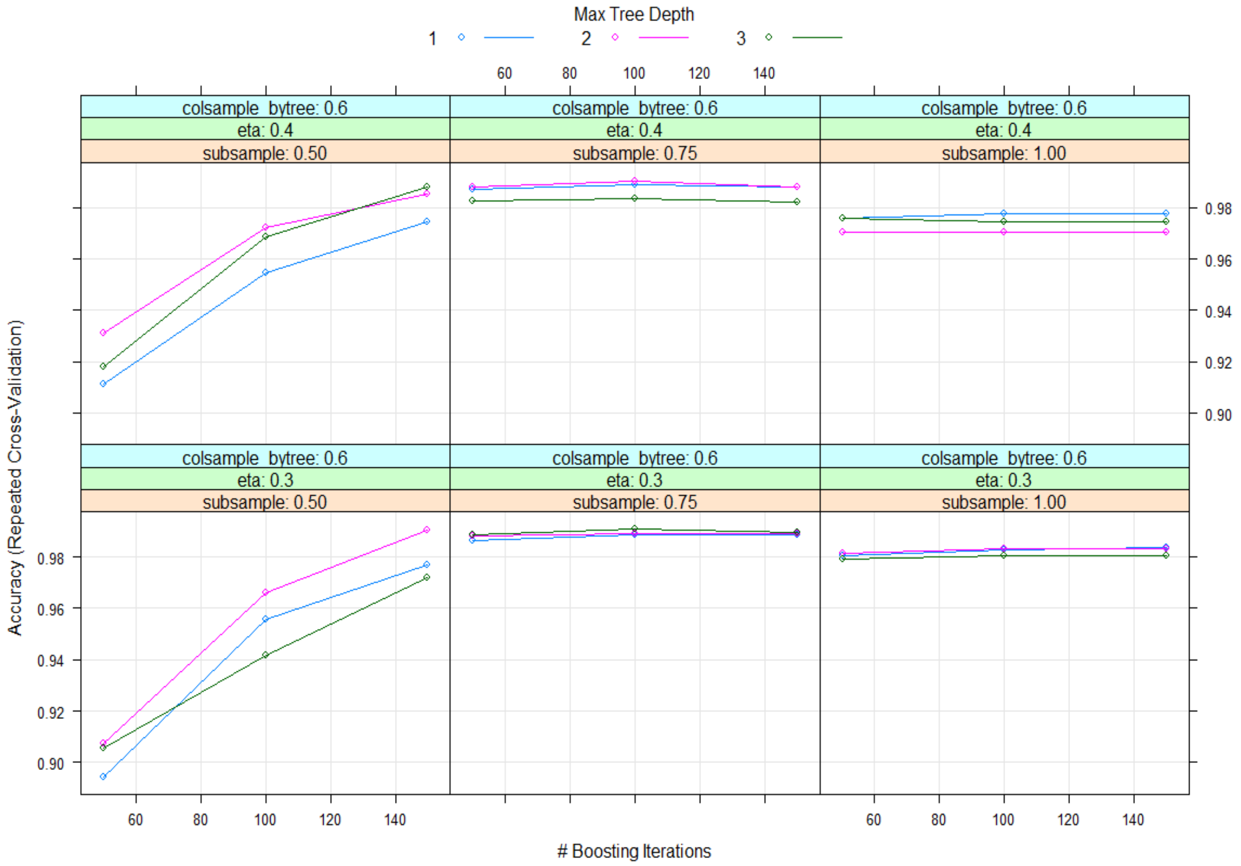


Figure 3.3: Cross-validation training curve for XGBoost

Similar to Random Forest, the top 20 important symptoms selected by XGBoost are as follows:

Table 3.3: The top 20 important symptoms selected by XGBoost

muscle_pain	**receiving_blood_transfusion**
fatigue	high_fever
chest_pain	diarrhea
mild_fever	**nausea**
skin_rash	**yellowing_of_eyes**
itching	yellowish_skin
excessive_hunger	**blood_in_sputum**
joint_pain	cough
vomiting	breathlessness
dark_urin	**weight_loss**

Among both Random Forest and XGBoost, the above bolded ten symptoms are treated as the crucial symptoms in determining the disease.

4. Results

In summary, using three different classification models – Decision Trees, Random Forest and XGBoost. The exact accuracy and kappa values for the selected hyperparameter values for all three models are as follows:

Table 4.1: Classification Models result

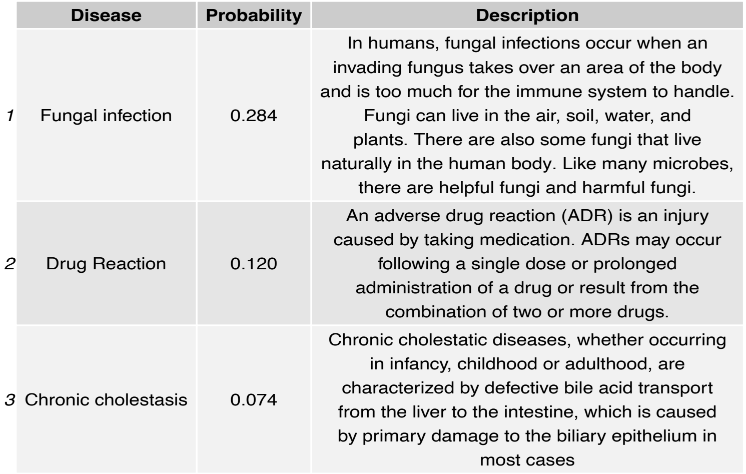
Model	Method	Package	Hyperparameter	Selection	Accuracy	Kappa
Decision Tree	ctree	Party	mincriterion	0.99	0.056	0.027
Random Forest	rf	randomForest	mtry	60	0.991	0.99
XGBoost	xgbTree	Xgboost, plyr	nrounds, max_depth, eta, colsample_bytree, subsample	100, 3, 0.3, 0.6, 0.75	0.99	0.99

If a model is to be chosen among the 3, then it would be Random Forest as it took less time to fit the data with exceptional accuracy.

5. Disease report generation

The main objective of the project is to model the variability of diseases based on symptoms and predict the disease of a patient based on the symptoms. This objective is achieved with ~99% accuracy by both **Random Forest** and **xgboost models**. But compared, Random Forest could be trained faster than the xgboost model. Thus, we considered Random Forest to be our final model.

So using the model, a report can be generated with the top 3 predicted diseases from a given set of symptoms, along with the probability and the description of that disease. While generating such a report, only known symptoms are accepted by the functionality. Thus, in case unknown symptoms are given, the code would stop and displays an error message. The symptoms are to be given in code for the report to be generated.



	Disease	Probability	Description
1	Fungal infection	0.284	In humans, fungal infections occur when an invading fungus takes over an area of the body and is too much for the immune system to handle. Fungi can live in the air, soil, water, and plants. There are also some fungi that live naturally in the human body. Like many microbes, there are helpful fungi and harmful fungi.
2	Drug Reaction	0.120	An adverse drug reaction (ADR) is an injury caused by taking medication. ADRs may occur following a single dose or prolonged administration of a drug or result from the combination of two or more drugs.
3	Chronic cholestasis	0.074	Chronic cholestatic diseases, whether occurring in infancy, childhood or adulthood, are characterized by defective bile acid transport from the liver to the intestine, which is caused by primary damage to the biliary epithelium in most cases

Figure 5.1: Top 3 Diseases PDF Report

6. Limitations

- Diseases present only in the dataset can be predicted; other diseases cannot be predicted.
- Not all symptoms of diseases are mentioned in the dataset, so if a new symptom that the dataset does not contain comes into account, then our model cannot predict the result.
- Dataset used is stagnant; the addition of a new disease or symptoms is to be done in the dataset and the model has to be trained again.